Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence

Le Noury, J.C.; Nardo, J.M.; Healy, David; Jureidini, J.; Raven, M.; Tufanaru, C.; Abi-Jaoude, E.

BMJ

DOI:
10.1136/bmj.h4320
10.1136/bmj.h4320

Published: 16/09/2015

Publisher's PDF, also known as Version of record

Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence

Joanna Le Noury,[1] John M Nardo,[2] David Healy,[1] Jon Jureidini,[3] Melissa Raven,[3] Catalin Tufanaru,[4] Elia Abi-Jaoude[5]

[1]School of Medical Sciences, Bangor University, Bangor, Wales, UK
[2]Emory University, Atlanta, Georgia, USA
[3]Critical and Ethical Mental Health Research Group, Robinson Research Institute, University of Adelaide, Adelaide, South Australia, Australia
[4]Joanna Briggs Institute, Faculty of Health Sciences, University of Adelaide, Adelaide, South Australia, Australia
[5]Department of Psychiatry, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada

Correspondence to: J Jureidini
Jon.Jureidini@adelaide.edu.au

Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmj.h4320)

Cite this as: *BMJ* 2015;351:h4320
doi: 10.1136/bmj.h4320

Accepted: 03 August 2015

ABSTRACT

OBJECTIVES

To reanalyse SmithKline Beecham's Study 329 (published by Keller and colleagues in 2001), the primary objective of which was to compare the efficacy and safety of paroxetine and imipramine with placebo in the treatment of adolescents with unipolar major depression. The reanalysis under the restoring invisible and abandoned trials (RIAT) initiative was done to see whether access to and reanalysis of a full dataset from a randomised controlled trial would have clinically relevant implications for evidence based medicine.

DESIGN

Double blind randomised placebo controlled trial.

SETTING

12 North American academic psychiatry centres, from 20 April 1994 to 15 February 1998.

PARTICIPANTS

275 adolescents with major depression of at least eight weeks in duration. Exclusion criteria included a range of comorbid psychiatric and medical disorders and suicidality.

INTERVENTIONS

Participants were randomised to eight weeks double blind treatment with paroxetine (20-40 mg), imipramine (200-300 mg), or placebo.

MAIN OUTCOME MEASURES

The prespecified primary efficacy variables were change from baseline to the end of the eight week acute treatment phase in total Hamilton depression scale (HAM-D) score and the proportion of responders (HAM-D score ≤8 or ≥50% reduction in baseline HAM-D) at acute endpoint. Prespecified secondary outcomes were changes from baseline to endpoint in depression items in K-SADS-L, clinical global impression, autonomous functioning checklist, self-perception profile, and sickness impact scale; predictors of response; and number of patients who relapse during the maintenance phase. Adverse experiences were to be compared primarily by using descriptive statistics. No coding dictionary was prespecified.

RESULTS

The efficacy of paroxetine and imipramine was not statistically or clinically significantly different from placebo for any prespecified primary or secondary efficacy outcome. HAM-D scores decreased by 10.7 (least squares mean) (95% confidence interval 9.1 to 12.3), 9.0 (7.4 to 10.5), and 9.1 (7.5 to 10.7) points, respectively, for the paroxetine, imipramine and placebo groups (P=0.20). There were clinically significant increases in harms, including suicidal ideation and behaviour and other serious adverse events in the paroxetine group and cardiovascular problems in the imipramine group.

CONCLUSIONS

Neither paroxetine nor high dose imipramine showed efficacy for major depression in adolescents, and there was an increase in harms with both drugs. Access to primary data from trials has important implications for both clinical practice and research, including that published conclusions about efficacy and safety should not be read as authoritative. The reanalysis of Study 329 illustrates the necessity of making primary trial data and protocols available to increase the rigour of the evidence base.

WHAT IS ALREADY KNOWN ON THIS TOPIC

There is a lack of access to data from most clinical randomised controlled trials, making it difficult to detect biased reporting

In the absence of access to primary data, misleading conclusions in publications of those trials can seem definitive

SmithKline Beecham's Study 329, an influential trial that reported that paroxetine was safe and effective for adolescents, is one such study

WHAT THIS STUDY ADDS

On the basis of access to the original data from Study 329, we report a reanalysis that concludes that paroxetine was ineffective and unsafe in this study

Access to primary data makes clear the many ways in which data can be analysed and represented, showing the importance of access to data and the value of reanalysis of trials

There are important implications for clinical practice, research, regulation of trials, licensing of drugs, and the sociology and philosophy of science

Our reanalysis required development of methods that could be adapted for future reanalyses of randomised controlled trials

Introduction

In 2013, in the face of the selective reporting of outcomes of randomised controlled trials, an international group of researchers called on funders and investigators of abandoned (unpublished) or misreported trials to publish undisclosed outcomes or correct misleading publications.[1] This initiative was called "restoring invisible and abandoned trials" (RIAT). The researchers identified many trials requiring restoration and emailed the funders, asking them to signal their intention to publish the unpublished trials or publish corrected versions of misreported trials. If funders and investigators failed to undertake to correct a trial that had been identified as unpublished or misreported, independent groups were encouraged to publish an accurate representation of the clinical trial based on the relevant regulatory information.

The current article represents a RIAT publication of Study 329. The original study was funded by

SmithKline Beecham (SKB; subsequently GlaxoSmithKline, GSK). We acknowledge the work of the original investigators. This double blinded randomised controlled trial to evaluate the efficacy and safety of paroxetine and imipramine compared with placebo for adolescents diagnosed with major depression was reported in the *Journal of the American Academy of Child and Adolescent Psychiatry (JAACAP)* in 2001, with Martin Keller as the primary author.[2] The RIAT researchers identified Study 329 as an example of a misreported trial in need of restoration. The article by Keller and colleagues, which was largely ghostwritten,[3] claimed efficacy and safety for paroxetine that was at odds with the data.[4] This is problematic because the article has been influential in the literature supporting the use of antidepressants in adolescents.[5]

On 14 June 2013, the RIAT researchers asked GSK whether it had any intention to restore any of the trials it sponsored, including Study 329. GSK did not signal any intent to publish a corrected version of any of its trials. In later correspondence, GSK stated that the study by Keller and colleagues "accurately reflects the honestly-held views of the clinical investigator authors" and that GSK did "not agree that the article is false, fraudulent or misleading."[6]

Study 329 was a multicentre eight week double blind randomised controlled trial (acute phase), followed by a six month continuation phase. SKB's stated primary objective was to examine the efficacy and safety of imipramine and paroxetine compared with placebo in the treatment of adolescents with unipolar major depression. Secondary objectives were to identify predictors of treatment outcomes across clinical subtypes; to provide information on the safety profile of paroxetine and imipramine when these drugs were given for "an extended period of time"; and to estimate the rate of relapse among patients who responded to imipramine, paroxetine, and placebo and were maintained on treatment. Study enrolment took place between April 1994 and March 1997.

The first RIAT trial publication was a surgery trial that had been only partly published before.[7] Few previously published randomised controlled trials have ever been subsequently reported in published papers by different teams of authors.[8]

Box 1 Study eligibility criteria

Inclusion criteria

- Adolescents aged 12-18 who met DSM-III-R criteria for major depression for at least 8 weeks
- Severity score <60 on the children's global assessment scale (CGAS)
- Score ≥12 on the Hamilton depression scale (17 item) (HAM-D)
- Medically healthy
- IQ ≥80 (based on Peabody picture vocabulary test)

Exclusion criteria

- Current or past DSM-III-R diagnosis of bipolar disorder, schizoaffective disorder, anorexia nervosa, bulimia, alcohol or drug abuse/dependence, obsessive-compulsive disorder, autism/pervasive mental disorder, or organic psychiatric disorder
- Current (within 12 months) DSM-III-R diagnosis of post-traumatic stress disorder
- Adequate trial of an antidepressant within six months (at least four weeks' treatment with an adequate dose of antidepressant)
- Suicidal ideation with a definite plan, suicide attempt during current depressive episode, or history of suicide attempt by drug overdose
- Medical illness that contraindicates the use of heterocyclic antidepressants
- Current use of psychotropic drugs (including anxiolytics, antipsychotics, mood stabilisers), or illicit drugs
- Organic brain disease, epilepsy, or "mental retardation"
- Pregnancy or lactation
- Sexually active females not using reliable contraception
- Use of an investigational drug within previous 30 days or five half lives of the investigation drug

Methods

We reanalysed the data from Study 329 according to the RIAT recommendations. To this end, we used the clinical study report (SKB's "final clinical report"), including appendices A-G, which are publically available on the GSK website,[9] other publically available documents,[10] and the individual participant data accessed through SAS Solutions OnDemand website,[11] on which GSK subsequently also posted some Study 329 documents (available only to users approved by GSK). After negotiation,[12] GSK posted about 77 000 pages of de-identified individual case report forms (appendix H) on that website. We used a tool for documenting the transformation from regulatory documents to journal publication, based on the CONSORT 2010 checklist of information to include when reporting a randomised trial. The audit record, including a table of sources of data consulted in preparing each part of this paper, is available in appendix 1.

Except where indicated, in accordance with RIAT recommendations, our methods are those set out in the 1994-96 protocol for Study 329.[13] In cases when the methods used and published by Keller and colleagues diverged from the protocol, we followed the original protocol. Because the protocol specified method of correction for missing values—last observation carried forward—has been questioned in the intervening years, we also included a more modern method—multiple imputation—at the request of the *BMJ* peer reviewers. This is a post hoc method added for comparison only and is not part of our formal reanalysis. When the protocol was not specific, we chose by consensus standard methods that best presented the data. The original 1993 protocol had minor amendments in 1994 and 1996 (replacement of the Schedule for Affective Disorders and Schizophrenia for Adolescents-Present Version with the Lifetime Version (K-SADS-L) and reduction in required sample size). Furthermore, the clinical study report (CSR) reported some procedures that varied from those specified in the protocol. We have noted variations that we considered relevant.

Participants

The original study recruited 275 adolescents aged 12-18 who met DSM-IV criteria[14] for a current episode of major depression of at least eight weeks' duration (the protocol specified DSM-III-R criteria, which are similar). Box 1 lists the eligibility criteria.

An unknown number of patients (not disclosed in the available documents) identified by telephone screening as potential participants were subsequently evaluated

doi: 10.1136/bmj.h4320 | *BMJ* 2015;351:h4320 | the bmj

at the study site by a senior clinician (psychiatrist or psychologist). Multiple meetings and teleconferences were held by the sponsoring company with site study investigators to ensure standardisation across sites. Patients and parents were interviewed separately with the K-SADS-L. After this initial assessment, the patient and parent both signed the study informed consent form; there was no mention of a separate assent form in the protocol or in the CSR. A screening period of seven to ten days was used to obtain past clinical records and to document that the depressive symptoms were stable. At the end of the screening period, only patients continuing to meet the inclusion criteria (DSM-III-R major depression and the HAM-D total score ≥12) were randomised. There was no placebo lead-in phase.

There were originally six study sites, but this was increased to 12 (10 in the United States and two in Canada). The centres were affiliated with either a university or a hospital psychiatry department and had experience with adolescent patients. The investigators were selected for their interest in the study and their ability to recruit study patients.

The recruitment period ran from 20 April 1994 until 15 March 1997, and the acute phase was completed on 7 May 1997. In a small number of patients, 30 day follow-up data for cases that went into the continuation phase were collected into February 1998.

Patient involvement

So far as we can ascertain, there was no patient involvement in SKB's study design.

Interventions

The study drug was provided to patients in weekly blister packs. Patients were instructed to take the drug twice daily. There were six dosing levels. Over the first four weeks, all patients were titrated to level four, corresponding to 20 mg paroxetine or 200 mg imipramine, regardless of response. Non-responders (those failing to reach responder criteria) could be titrated up to level five or six over the next four weeks. This corresponds to maximum doses of 60 mg paroxetine and 300 mg imipramine.

Compliance with treatment was evaluated from the number of capsules dispensed, taken, and returned. Non-compliance was defined as taking less than 80% or more than 120% of the number of capsules, assessed from the numbers expected to be returned at two consecutive visits, and resulted in withdrawal. Any patient missing two consecutive visits was also withdrawn from the study.

Patients were provided with 45 minute weekly sessions of supportive psychotherapy,[15] primarily for the purpose of assessing the effects of treatment.

Sample size

The acute phase of the trial was initially based on a power analysis that indicated that a sample size of 100 patients per treatment group was required to have a statistical power of 80% for a two tailed α of 0.05 and an effect size of 0.40. This effect size entailed a difference of 4 in the HAM-D total score from baseline to endpoint, specified in the protocol to be large enough to be clinically meaningful, considering a standard deviation of 10. No allowance was made in the power calculation for attrition (anticipated dropout rate) or non-compliance during the study.

Recruitment was slower than expected, and reportedly supplies of treatment (mainly placebo) ran short due to exceeding the expiry date. The researchers carried out a midcourse evaluation of 189 patients, without breaking the blinding, which showed less variability in HAM-D scores (SD 8) than expected. Therefore the recruitment target was reduced to 275 on the grounds that it would have no negative impact on the estimated 80% power required to detect a 4 point difference between placebo and active drug groups.

Randomisation

A computer generated randomisation list of 360 numbers for the acute phase was generated and held by SKB. According to the CSR, treatments were balanced in blocks of six consecutive patients; however, there is an inconsistency in that appendix A randomisation code details block sizes of both six and eight. Each investigator was allocated a block of consecutively numbered treatment packs, and patients were assigned treatment numbers in strict sequential order. Patients were randomised in a 1:1:1 ratio to treatment with paroxetine, imipramine, or placebo.

Blinding

Paroxetine was supplied as film coated, capsule shaped yellow (10 mg) and pink (20 mg) tablets. Imipramine (50 mg) was bought commercially and supplied as green film coated round 50 mg tablets. "Paroxetine placebos" matched the paroxetine 20 mg tablets, and "imipramine placebos" matched the imipramine tablets. All tablets were over-encapsulated in bluish-green capsules to preserve blinding.

The blinding was to be broken only in the event of a serious adverse event that the investigator thought could not be adequately treated without knowing the identity of the allocated study treatment. The identity of the study treatment was not otherwise to be disclosed to the investigator or SKB staff associated with the study.

Outcomes

Patients were evaluated weekly for the following outcome variables during the eight week duration of the acute treatment phase.

Primary efficacy variables

The prespecified primary efficacy variables were change in total score on HAM-D[16] from the beginning of the treatment phase to the endpoint of the acute phase and the proportion of responders at the end of the eight week acute treatment phase (longer than many antidepressant trials). Responders were defined as patients who had ≥50% reduction in the HAM-D or a HAM-D score of ≤8. (Scores on the HAM-D can vary from 0 to 52.)

Secondary efficacy variables

The prespecified secondary efficacy variables were:

- Changes from baseline to endpoint in:
  - Depression items in K-SADS-L
  - Clinical global impression (CGI)
  - Autonomous functioning checklist[17]
  - Self perception profile
  - Sickness impact scale.
- Predictors of response (endogenous subtypes, age, previous episodes, duration and severity of present episode, comorbidity with separate anxiety, attention deficit, and conduct disorder)
- The number of patients who relapsed during the maintenance phase (referred to in the CSR and in this paper as "continuation phase").

Both before and after breaking the blind, however, the sponsors made changes to the secondary outcomes as previously detailed.[4] We could not find any document that provided any scientific rationale for these post hoc changes,[18] and the outcomes are therefore not reported in this paper.

Challenges in carrying out RIAT

To our knowledge this is the first RIAT analysis of a misreported trial by an external team of authors, so there are no clear precedents or guides. Challenges we have encountered included:

Potential or perceived bias

A RIAT report is not intended to be a critique of a previous publication. The point is rather to produce a thorough independent analysis of a trial that has remained unpublished or called into question. We acknowledge, however, that any RIAT team might be seen as having an intrinsic bias in that questioning the earlier published conclusions is what brought some members of the team together. Consequently, we took all appropriate procedural steps to avoid such putative bias. In addition, we have made the data available for others to analyse.

Correction for testing multiple variables

We had multiple sources of information: the protocol; the published paper; the documents posted on the GSK website including the CSR and individual patient data; and the raw primary data in the case report forms provided by GSK on a remote desktop for this project. The protocol declared two primary and six secondary variables for the three treatment groups in two differing datasets (observed case and last observation carried forward). The CSR contained statistical comparisons on 28 discrete variables using two comparisons (paroxetine *v* placebo and imipramine *v* placebo) in the two datasets (observed case and last observation carried forward). The published paper listed eight variables with two statistical comparisons each in one dataset (last observation carried forward). The authors of the original paper, however, did not deal with the need for corrections for multiple variables—a standard requirement when there are multiple outcome measures. In the final analysis, there were no statistically or clinically significant findings for any outcome variable, so corrections were not needed for this analysis.

Statistical testing

The protocol called for ANOVA testing (generalised linear model) for continuous variables using a model that included the effects of site, treatment, and site × treatment interaction, with the latter dropped if $P \geq 0.10$. Logistical regression (2×3 χ^2) was prescribed for categorical variables under the same model. Both methods begin with an omnibus statistic for the overall significance of the dataset, then progress to pairwise testing if, and only if, the omnibus statistic meets $\alpha = 0.05$. Yet all statistical outcomes in the CSR and published paper were reported only as the pairwise values for only two of the three possible comparisons (paroxetine *v* placebo and imipramine *v* placebo), with no mention of the omnibus statistic. Therefore, we conducted the required omnibus analyses, with negative results as shown. The pairwise values are available in table A in appendix 2.

Missing values

The protocol called for evaluation of the observed case and last observation carried forward datasets, with the latter being definitive. The last observation carried forward method for correcting missing values was the standard at the time the study was conducted. It continues to be widely used, although newer models such as multiple imputation or mixed models are superior. We chose to adhere to the protocol and use the last observation carried forward method, including multiple imputation for comparison only.

Outcome variables not specified in protocol

There were four outcome variables in the CSR and in the published paper that were not specified in the protocol. These were the only outcome measures reported as significant. They were not included in any version of the protocol as amendments (despite other amendments), nor were they submitted to the institutional review board. The CSR (section 3.9.1) states they were part of an "analysis plan" developed some two months before the blinding was broken. No such plan appears in the CSR, and we have no contemporaneous documentation of that claim, despite having repeatedly requested it from GSK.

Conclusions

We decided that the best and most unbiased course of action was to analyse the efficacy data in the individual patient data based on the last guaranteed a priori version of SKB's own protocol (1994, amended in 1996 to accept a reduced sample size). Although the protocol omitted a discussion of corrections that we would have thought necessary, correction for multiple variables is designed to prevent false positives and there were no positives. We agreed with the statistical mandates of the protocol, but though we regarded pairwise comparisons in the absence of overall significance as inappropriate, we recognise that this is not a universal opinion, so we included the data in table A in appendix 2.

doi: 10.1136/bmj.h4320 | *BMJ* 2015;351:h4320 | the bmj

Finally, although investigators can explore the data however they want, additional outcome variables outside those in the protocol cannot be legitimately declared once the study is underway, except as "exploratory variables"—appropriate for the discussion or as material for further study but not for the main analysis. The a priori protocol and blinding are the bedrock of a randomised controlled trial, guaranteeing that there is not even the possibility of the HARK phenomenon ("hypothesis after results known"). Though we can readily show that none of the reportedly "positive" four non-protocol outcome variables stands up to scrutiny, the primary mandate of the RIAT enterprise is to reaffirm essential practices in randomised controlled trials, so we did not include these variables in our efficacy analysis.

Harm endpoints

An adverse experience/event was defined in the protocol (page 18) as "any noxious, pathologic or unintended change in anatomical, physiologic or metabolic functions as indicated by physical signs, symptoms and/or laboratory changes occurring in any phase of the clinical trial whether associated with drug or placebo and whether or not considered drug related. This includes an exacerbation of pre-existing conditions or events, intercurrent illnesses, drug interaction or the significant worsening of the disease under investigation that is not recorded elsewhere in the case report form under specific efficacy assessments."

Adverse events were to be elicited by the investigator asking a non-leading question such as: "Do you feel different in any way since starting the new treatment/the last assessment?" Details of adverse events that emerged with treatment, their severity, including any change in study drug administration, investigator attribution to study drug, any corrective therapy given, and outcome status were documented. Attribution or relation to study drug was judged by the investigator to be "unrelated," "probably unrelated," "possibly related," "probably related," or "related."

Vital signs and electrocardiograms were obtained at weekly visits. Patients with potentially concerning cardiovascular measures either had their drug dose reduced or were withdrawn from the study. In addition, if the combined serum concentrations (obtained at weeks four and eight) of imipramine and desipramine exceeded 500 μg/mL the patient was to be withdrawn from the study.

Clinical laboratory tests, including clinical chemistry, haematology, and urinalysis, were carried out at the screening visit and at the end of week eight. Clinically relevant laboratory abnormalities were to be included as adverse events.

Source of harms data

The harms data in this paper cover the acute phase, a taper period, and a follow-up phase of up to 30 days for those who discontinued treatment because of adverse events. To ensure comparability with the report by Keller and colleagues, none of the tables contains data from the continuation phase.

Data on adverse events come from the CSR lodged on GSK's website,[19] primarily in appendix D. Appendix B provides details of concomitant drugs. Additional information was available from the summary narratives in the body of the CSR for patients who had adverse events that were designated as serious or led to withdrawal. (Of the 11 patients taking paroxetine who experienced adverse events designated as serious, nine discontinued treatment because of these events.) The large number of other patients discontinued because of adverse events that were not regarded as serious, or discontinued for lack of efficacy or protocol violations, however, did not generate patient narratives. The tables in appendix D of the CSR provide the verbatim terms used by the blinded investigators, along with preferred terms as coded by SKB using the adverse drug events coding system (ADECS) dictionary. Appendix D also includes ratings of severity and ratings of relatedness. We used the Medical Dictionary for Regulatory Activities (MedDRA) to code the verbatim terms provided in appendix D in the CSR. MedDRA terminology is the international medical terminology developed under the auspices of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) www.meddra.org), endorsed by the FDA, and now used by GSK.[20]

Several limitations of the ADECS coded preferred terms provided in appendix D of the CSR became clear when we examined the ADECS preferred terms assigned to the verbatim terms. Firstly, several verbatim terms had been left uncoded into ADECS. Secondly, several adverse events found in the patient narratives of serious adverse events that led to discontinuation from the trial were not transcribed into appendix D.

We therefore approached GSK for access to case report forms (appendix H of the CSR), which are not publically available. GSK made available all 275 case report forms for patients entered into Study 329. These forms, however, which totalled about 77 000 pages, were available only through a remote desktop facility (SAS Solutions OnDemand Secure Portal),[11] which made it difficult and extremely time consuming to inspect the records properly.[21] Effectively only one person could undertake the task, with backup for ambiguous cases. Accordingly we could not examine all case report forms. Instead we decided to focus on those 85 participants identified in appendices D and G of the CSR who were withdrawn from the study, along with eight further participants who were known from our inspection of the CSRs to have become suicidal. Of the case report forms that were checked, 31 were from the paroxetine group, 40 from the imipramine group, and 22 from the placebo group.

All case report forms were reviewed by JLN, who was trained in the use of MedDRA. The second reviewer (JMN), a clinician, was not trained in the MedDRA system, but training is not necessary for coding of dropouts. These two reviewers agreed about reasons for discontinuation and coding of side effects (we did not use a quantitative indicator of agreement between raters). We scrutinised these 93 case report forms for all

adverse events occurring during the acute, taper, and follow-up phases, and compared our totals for adverse events with the totals reported in appendix D of the CSR. This review process identified additional adverse events that had not been recorded as verbatim terms in appendix D of the CSR. It also led to recoding of several of the reasons for discontinuation. Tables B, C, and H in appendix 2 show the new adverse events and the reasons for changing the discontinuation category.

At least 1000 pages were missing from the case report forms we reviewed, with no discernible pattern to missing information—for example, one form came with a page inserted stating that pages 114 to 223 were missing, without indicating reasons.

Coding of adverse events

Choice of coding dictionary for harms

The protocol (page 25) indicates that adverse events were to be coded and compared by preferred term and body system by using descriptive statistics but does not prespecify a choice of coding dictionary for generating preferred terms from verbatim terms. The CSR (written after the study ended) specifies that the adverse events noted by clinical investigators in this trial were coded with ADECS, which was being used by SKB at the time. This system was derived from a coding system developed by the US Food and Drug Administration (FDA), Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART), but ADECS is not itself a recognised system and is no longer available.

We coded adverse events using MedDRA, which has replaced COSTART for the FDA because it is by far the most commonly used coding system today. For coding purposes, we have taken the original terms used by the clinical investigators, as transcribed into appendix D of the CSR, and applied MedDRA codes to these descriptions. Information from appendix D was transcribed into spreadsheets (available at www.Study329.org). The verbatim terms and the ADECS coding terms were transcribed first into these sheets, allowing all coding to be done before the drug names were added in. The transcription was carried out by a research assistant who was a MedDRA trained coder but took no part in the actual coding. All coding was carried out by JLN, and checked by DH, or vice versa. All of our coding from the verbatim terms in the appendix D of the CSR was done blind, as was coding from the case report forms.

We present results as SKB presented them in the CSR using the ADECS dictionary (table 14.2.1), and as coded by us using MedDRA. In general, MedDRA coding stays closer than ADECS to the original clinician description of the event. For instance, MedDRA codes "sore throat" as "sore throat" but SKB, using ADECS, coded it as "pharyngitis" (inflammation of the throat). Sore throats can arise because of pharyngitis, but when someone is taking selective serotonin reuptake inhibitors they can indicate a dystonic reaction in the oropharyngeal area.[22]

Classification of a problem as a "respiratory system disorder" (inflammation) rather than as a "dystonia" (a central nervous system disorder) can make a considerable difference to the apparent adverse event profile of a drug. In staying closer to the original description of events, MedDRA codes suicidal events as "suicidal ideation" or "self harm/attempted suicide" rather than the ADECS option of "emotional lability"; similarly, aggression is more clearly flagged as "aggressive events" rather than "hostility."

Most coding was straightforward. Nearly all the verbatim terms simply mapped onto coding terms in MedDRA. Coding challenges usually related to cases where there were significant adverse events but the patients were designated by SKB to have discontinued for lack of efficacy. There was no patient narrative for such patients, in contrast to patients deemed to have discontinued because of the adverse event occurring at discontinuation. There were few challenging coding decisions. Appendix 3 shows our coding of cases in which suicidal and self injurious behaviours were considered.

Analysis of harms data

In analysing the harms data for the safety population, we firstly explored the discrepancies in the number of events between case report forms and the CSR. Secondly, we presented all adverse events rather than those happening only at a particular rate (as done by Keller and colleagues). Thirdly, we grouped events into broader system organ class (SOC) groups: psychiatric, cardiovascular, gastrointestinal, respiratory, and other. Table D in appendix 2 summarises all adverse events by all MedDRA SOC groupings. Fourthly, we broke down events by severity, selecting adverse events coded as severe and using the listing in appendix G of the CSR of patients who discontinued for any reason. Fifthly, we included an analysis of the effects of previous treatment, presenting the run-in phase profiles of drugs taken by patients entering each of the three arms of the study and comparing the list of adverse events experienced by patients on concomitant drugs (from appendix B) versus those not on other drugs. Finally, we extracted the events occurring during the taper and follow-up phase.

We did not undertake statistical tests of harms data, as discussed below.

Patient withdrawal

A study patient could withdraw or be withdrawn prematurely for "adverse experiences including intercurrent illness," "insufficient therapeutic effect," "deviation from protocol including non-compliance," "loss to follow-up," "termination by SB [SKB]," and "other (specify)."

The CSR states that the primary reason for withdrawal was determined by the investigator. We reviewed the codes given for discontinuation from the study, which are found in appendix G of the CSR, and we made changes in a proportion of cases.

Statistical methods

The primary population of interest was the intention to treat population that included all patients who received at least one dose of study drug and had at least one assessment of efficacy after baseline. The demographic characteristics, description of the baseline depressive

episode, additional psychiatric diagnoses, and personal history variables of the patients were summarised descriptively by treatment group.

The acute phase eight week endpoint was our primary interest. Statistical conclusions concerning the efficacy of paroxetine and imipramine were made by using data obtained from the last observation carried forward (that is, the last assessment "on therapy" during the acute phase) and observed case datasets. Paroxetine and imipramine were each to be compared with placebo; there was to be no comparison of paroxetine with imipramine.

We followed the methods of the a priori 1994 study protocol (amended in 1996 to accept a reduced sample size). It did not provide explicit statistical hypotheses (null hypotheses and alternative hypotheses); nor were there justifications for the proposed statistical approaches or statistical assumptions underlying them.

One of the two primary efficacy variables, proportion of responders (response), and one secondary efficacy variable, proportion of patients relapsing, were treated as categorical variables. The second primary efficacy variable, change in total HAM-D score over the acute phase, and the remaining secondary efficacy variables were treated as continuous variables.

In accordance with the protocol, the continuous variables were analysed with parametric analysis of variance (ANOVA) with effects in the model including treatment, investigator, and treatment by investigator interaction. Pairwise comparisons were not done if the omnibus (overall) ANOVA was not significant (two sided P<0.05), as specified by the protocol (we acknowledge differing opinions about this issue in the statistical literature,[23] so we included them in table A in appendix 2, for completeness). The categorical variables were analysed with logistic regression, with the same effects included. In either case, if the treatment by investigator interaction resulted in a two sided P>0.10, the interaction term was dropped from the model. Statistical testing was done with the linear model (LM) and general linear models (GLM) procedures of the R statistical package (version 2.15.2) as provided by GSK. Imputation was performed with the multiple imputation by chained equations (MICE) package also in R.[24]

For the analyses of relapse rates, we included all responders (HAM-D ≤8 or ≥50% reduction in symptoms) who met the original criteria for entry to the continuation phase of the study. Patients were considered to have relapsed if they no longer met the responder criteria or if they were withdrawn for "intentional overdose."

Table 1 | Baseline characteristics of groups in Study 329

	Paroxetine (n=93)	Imipramine (n=95)	Placebo (n=87)
Mean (SD) age (years)	14.8 (1.6)	14.9 (1.6)	15.1 (1.6)
Sex (male/female)	35/58	39/56	30/57
No (%) by race:			
White	77 (83)	83 (87)	70 (81)
African American	5 (5)	3 (3)	6 (7)
Asian American	1 (1)	2 (2)	2 (2)
Other	10 (11)	7 (7)	9 (10)
Depression:			
Mean (SD) duration of episode (months)	14 (18)	13 (17)	13 (17)
Mean (SD) age at first episode (years)	13.1 (2.8)	13.7 (2.7)	13.5 (2.3)
No (%) of previous episodes:			
0	0 (0)	2 (2)	0 (0)
1	75 (81)	75 (79)	68 (77)
2	11 (12)	13 (14)	12 (14)
>3	7 (7)	5 (6)	7 (8)
No (%) with comorbidity:			
Any comorbid disorder *	42 (41)	47 (50)	39 (41)
Current anxiety disorder*	24 (19)	24 (26)	24 (19)
ODD, CD, or ADHD*	23 (25)	24 (26)	17 (20)
Least squares mean baseline scores (SEM):			
HAM-D	18.9 (0.44)	18.1 (0.43)	19.0 (0.44)
K-SADS-L	28.3 (9.5)	27.5 (0.51)	28.3 (0.52)
Autonomous function	93.4 (3.1)	97.0 (3.1)	94.2 (3.2)
Self perception profile	64.0 (2.2)	63.5 (2.2)	63.4 (2.3)
Sickness impact profile	32.4 (1.2)	30.8 (1.2)	32.9 (1.3)

ODD=oppositional defiant disorder, CD=conduct disorder, ADHD=attention-deficit/hyperactivity disorder, HAM-D=Hamilton depression scale, K-SADS-L=affective disorders and schizophrenia for adolescents-lifetime version, SD=standard deviation, SEM=standard error of mean.
*From K-SADS-L structured interview at screening.

Results

Table 1 shows the demographics of the groups, along with depression parameters, comorbidities, and baseline scores for the efficacy variables.

Figure 1 summarises the allocations and discontinuations among the three treatment groups during the acute study period. The flow chart covers the intention to treat population for the acute phase and the efficacy analysis. The paroxetine group was titrated to a dose of 20 mg/day by week four, with 55% (51/93) of participants moving to a higher dose (mean 28.0 mg/day, SD 8.4 mg) by week eight. The imipramine group was titrated to 200 mg/day by week four, with 40% (38/95) moving to a higher dose (mean 205.8 mg/day, SD 63.9 mg) by week eight. Twenty eight patients reached the highest permissible dose of 40 mg of paroxetine, and 20 patients were titrated to the maximum 300 mg of imipramine.

Efficacy

There were no discrepancies between any of our analyses and those contained in the CSR. Figures 2 and 3 illustrate the longitudinal values for the two primary efficacy variables: mean change from baseline in the HAM-D score and the percentage responding, defined as a decrease in HAM-D score by 50% or more from baseline or a final HAM-D score of ≤8. The difference between paroxetine and placebo fell short of the prespecified level of clinical significance (4 points) and neither primary outcome achieved significance at any measured interval for any dataset during the acute phase.

The formal reanalysis included both observed case and last observation carried forward datasets. As mentioned above, the multiple imputation dataset is included for comparison. There was no statistical significance (considered at P<0.05) or clinical significance shown for any of the prespecified primary or secondary efficacy variables in either the observed case or last observation carried forward datasets, so pairwise analysis was considered unjustified. Table 2 shows the results at week eight for reduction in HAM-D score and for the proportion of patients who met criteria for HAM-D response.

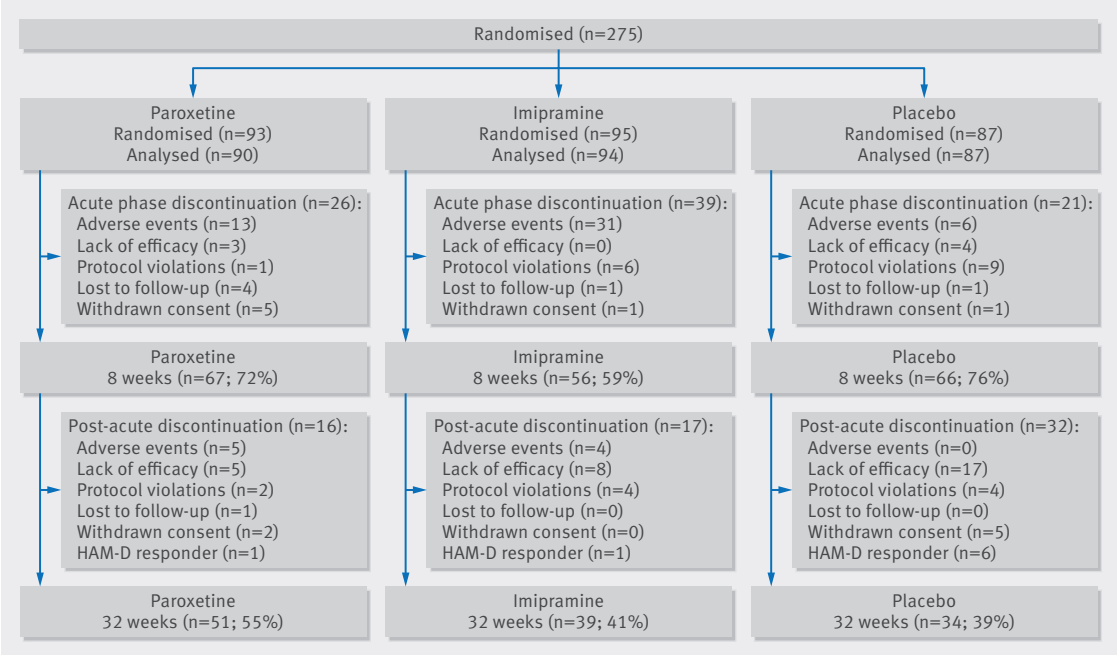

Fig 1 | Group allocations and discontinuations in trial of paroxetine and imipramine in treatment of major depression in adolescence

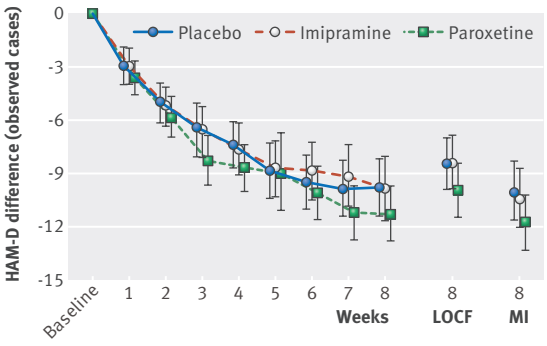

Fig 2 | Differences in HAM-D scores in study of efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence (table 2 shows numerical values). Points are least squares means (95% CI). LOCF=last observation carried forward, MI=multiple imputation

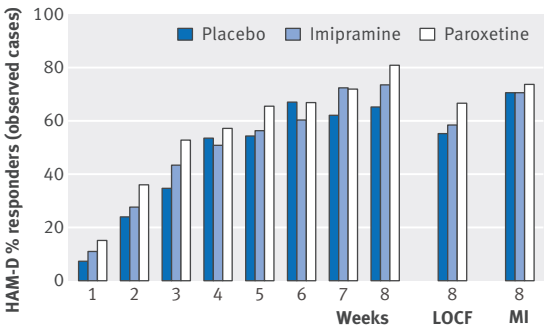

Fig 3 | Differences in HAM-D % responders in study of efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence (table 2 shows numerical values). LOCF=last observation carried forward, MI=multiple imputation

HAM-D scores decreased by 10.7 (95% confidence interval 9.1 to 12.3), 9.0 (7.4 to 10.5), and 9.1 (7.5 to, 10.7) points (least squares mean) for the paroxetine, imipramine, and placebo groups, respectively.

Table 3 shows the results at eight weeks for the secondary efficacy variables.

Although the protocol listed "predictors of response" among the secondary efficacy variables, the absence of statistically or clinically significant differences among the three arms rendered this analysis void.

The protocol also listed the relapse rate in the continuation phase for responders as a secondary outcome variable. Our calculation differed from that in the CSR because we included those whose HAM-D scores rose above the "response" range and those who intentionally overdosed. In the continuation phase, the dropout rates were too high in all groups for any precise interpretation: 33/51 (65%) in the paroxetine group; 25/39 (64%) in the imipramine group; and 21/34 (62%) in the placebo group. The recorded relapses were 25/51 (49%), 16/39 (41%), and 12/34 (35%), respectively. Although the relapse rate was lower in the placebo group, the differences were not significant (2×3 χ^2 P=0.44).

Harms

Review of case report forms

We reviewed case report forms in appendix H for 93 (34%) of 275 patients. We discovered adverse events recorded onto case report forms but not transcribed into the patient level listings of adverse events in appendix D of the CSR. Table 4 shows these discrepancies. The most common categories of additional adverse events found in case report forms were psychiatric for paroxetine (12/23) and placebo (4/10) and cardiovascular for imipramine (5/17) (table B in appendix 2).

doi: 10.1136/bmj.h4320 | *BMJ* 2015;351:h4320 | the bmj

Table 2 | Datasets for primary efficacy variables at eight weeks and proportion of patients who met criteria for HAM-D response >50% drop or <8 in Study 329 for observed cases (OC), last observation carried forward (LOCF), and multiple imputation

Data	Paroxetine		Imipramine		Placebo		P value
HAM-D change							
	Least squares mean (95% CI), SEM	No of patients	Least squares mean (95% CI), SEM	No of patients	Least squares mean (95% CI), SEM	No of patients	ANCOVA*
OC	–12.2 (–13.1 to –10.5), 0.88	67	–10.6 (–12.5 to –8.7), 0.97	56	–10.5 (–12.3 to –8.8), 0.88	66	0.26
LOCF	–10.7 (–12.3 to –9.1), 0.81	90	–9.0 (–10.5 to –7.4), 0.81	94	–9.1 (–10.7 to –7.5), 0.83	87	0.20
MI	–12.5 (–14.2 to –10.9), 0.83	90	–11.1 (–12.9 to –9.4), 0.89	94	–10.7 (–12.4 to –9.1), 0.83	87	0.24
HAM-D response (>50% reduction or <8)							
	Criteria met	Yes/no	Criteria met	Yes/no	Criteria met	Yes/no	$\chi 2$
OC	80.6%	54/13	73.2%	41/15	65.2%	43/23	0.13
LOCF	66.7%	60/30	58.5%	55/39	55.2%	48/39	0.27
MI	73.3%	66/24	70.2%	66/28	70.1%	61/26	0.24

HAM-D=Hamilton depression scale.
*All P values uncorrected for multiple variable sampling.

Table 3 | Datasets for secondary efficacy variables at eight weeks in Study 329 for observed cases (OC), last observation carried forward (LOCF), and multiple imputation

	Paroxetine			Imipramine			Placebo			
Data	Least squares mean (95% CI)	SEM	No of patients	Least squares mean (95% CI)	SEM	No of patients	Least squares mean (95% CI)	SEM	No of patients	P value*
K-SADS-L change										
OC	–12.1 (–13.8 to –10.3)	0.91	67	–10.7 (–12.7 to –8.7)	0.82	56	–10.7 (–12.5 to –8.9)	0.92	65	0.46
LOCF	–11.4 (–13.1 to –9.8)	0.84	83	–9.5 (–11.1 to –7.9)	0.82	88	–9.4 (–11.0 to –7.8)	0.83	85	0.13
MI	–12.3 (–13.9 to –10.6)	0.84	83	–11.5 (–13.3 to –9.7)	0.91	88	–10.9 (–12.6 to –9.2)	0.86	85	0.45
Clinical global impression mean score										
OC	1.9 (1.6 to 2.2)	0.15	68	2.2 (1.8 to 2.5)	0.17	56	2.4 (2.1 to 2.7)	0.16	66	0.09
LOCF	2.4 (2.1 to 2.7)	0.16	90	2.7 (2.4 to 3.0)	0.15	94	2.7 (2.4 to 3.0)	0.16	87	0.16
MI	1.9 (1.6 to 2.2)	0.14	90	2.2 (1.9 to 2.5)	0.15	94	2.4 (2.1 to 2.6)	0.14	87	0.07
Autonomous function check list change										
OC	14.4 (8.8 to 19.9)	2.83	58	13.3 (7.3 to 19.4)	3.04	52	9.3 (3.8 to 14.8)	2.81	60	0.32
LOCF	14.7 (9.2 to 20.2)	2.80	60	11.6 (5.8 to 17.3)	2.92	57	9.3 (8.1 to 17.2)	2.76	62	0.39
MI	14.0 (8.7 to 19.3)	2.65	60	14.5 (9.4 to 19.6)	2.60	57	9.1 (4.2 to 14.1)	2.52	62	0.24
Self perception profile change										
OC	12.9 (8.3 to 17.5)	2.31	60	13.2 (8.4 to 18.1)	2.46	55	12.7 (6.9 to 15.9)	2.30	60	0.88
LOCF	13.2 (8.6 to 17.8)	2.33	61	13.1 (8.3 to 17.8)	2.41	60	11.4 (6.9 to 15.9)	2.27	63	0.88
MI	15.4 (10.7 to 20.0)	2.35	61	14 (8.9 to 19.2)	2.60	60	14.7 (10.0 to 19.4)	2.39	63	0.92
Sickness impact profile change										
OC	–11.2 (–14.3 to –8.1)	1.57	62	–13.5 (–16.9 to –10.2)	1.70	55	–10.6 (–13.7 to –7.5)	1.57	62	0.24
LOCF	–11.4 (–14.4 to –8.3)	1.55	63	–13.0 (–16.2 to –9.8)	1.62	60	–9.9 (–12.9 to –6.9)	1.51	65	0.23
MI	–11.5 (–14.2 to –8.7)	1.39	63	–13.9 (–16.8 to –10.9)	1.50	60	–10.1 (–13.0 to –7.1)	1.48	65	0.19

K-SADS-L=affective disorders and schizophrenia for adolescents-lifetime version.
*ANCOVA. All P values uncorrected for multiple variable sampling.

Table 4 | Adverse events found in case report forms (CRFs) compared with adverse events listed in appendix D of clinical study report of Study 329

	Paroxetine (n=31)	Imipramine* (n=40)	Placebo (n=22)
Adverse events found in CRFs (appendix H)	159	257	77
Adverse events found in appendix D	136	240	67
% underestimate in relying only on appendix D	14%	7%	13%

*In considering adverse effects from imipramine, it should be noted that doses (mean 205.8 mg) were high for adolescents. In six comparator studies submitted by SKB as part of their 1991 approval NDA for paroxetine in adults, mean imipramine dose overall was 140 mg, with mean endpoint dose of 170 mg.[25]

Coding and representation of adverse event data

Table 5 presents the number of adverse events found in this study summarised by system organ class (SOC), firstly as coded by SKB using ADECS, secondly as reported by Keller and colleagues (who reported only adverse events that occurred at frequency of more than 5%), and thirdly as coded by us using MedDRA. Some adverse events always fall within a particular system organ class; others require that the coder choose between system organ classes. A full listing of adverse events can be found in table E in appendix 2.

We included events occurring during the taper phase that SKB allocated to the continuation phase as acute phase adverse events. In a study that has a continuation phase, the assessment of adverse events throws up a methodological difficulty not yet addressed by groups such as CONSORT. If a study has only an acute phase, then all adverse events are counted for all patients receiving treatment as well as in any taper phase, and often for a 30 day follow-up period. When a study has a continuation phase, the taper and 30 day follow-up periods are displaced. To ensure comparable analysis of all participants, we tallied the adverse events across the acute phase and both taper and follow-up phases, whether displaced or

Table 5 | Adverse events in SKB clinical study report (CSR) (ADECS coded), Keller and colleagues (ADECS coded), and RIAT reanalysis (MedDRA coded) in Study 329

	Paroxetine (n=93)			Imipramine (n=95)			Placebo (n=87)		
Adverse event (system organ class)	CSR*	Keller*	RIAT†	CSR*	Keller*	RIAT†	CSR*	Keller*	RIAT†
Cardiovascular	7	5	44	60	42	130	12	6	32
Gastrointestinal/digestive	80	84	112	108	106	147	59	61	79
Psychiatric	—	—	103	—	—	63	—	—	24
Respiratory	39	33	42	32	27	22	43	37	39
Neurological/nervous system	106	115	101	117	135	114	42	65	77
Other	121	28	79	51	30	76	30	38	79
Body as whole	106	—	—	125	—	—	121	—	—
Total	338	265	481	493	340	552	277	207	330

*Coded with ADECS (adverse drug events coding system). While in CSR (table 14.2.1—it is not clear whether this includes taper phase), headaches were included in "body as whole"; in paper by Keller and colleagues, adverse events "headache" and "dizziness" were grouped with psychiatric adverse events under heading "nervous system."
†Coded with MedDRA. MedDRA allows dizziness to be coded under "cardiovascular" or "neurological" SOCs and puts headaches under "neurological" SOC. See also tables D and E in appendix 2.

not. SKB do not seem to have done this, leading to some differences in numbers.

Figure 4 shows when suicidal and self injurious events occurred.

Table 6 shows the numbers of suicidal and self-injurious behaviours that we identified in our RIAT analysis and compared with what was reported by Keller and colleagues and documented in the CSR (table 6).

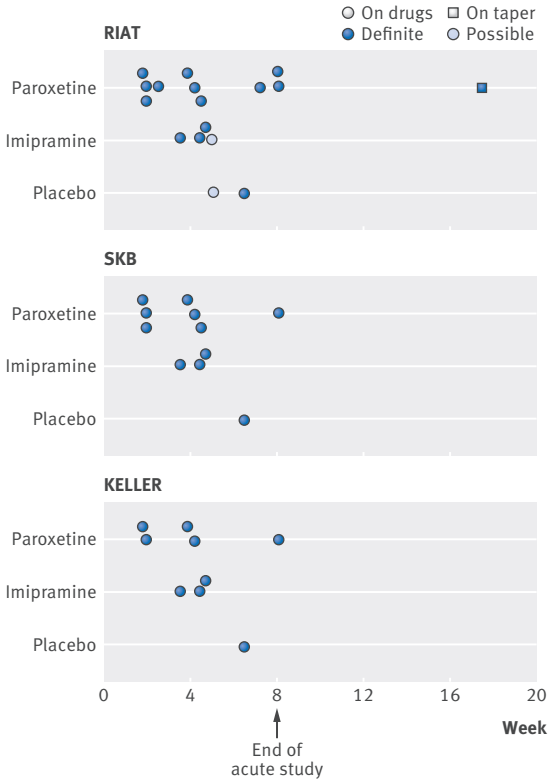

Fig 4 | Timing of suicidal and self injurious events in Study 329, Keller and colleagues, and RIAT analysis

Table 6 | Numbers of patients with suicidal and self injurious behaviours in Study 329 with different safety methods

	Paroxetine (n=93)	Imipramine (n=95)	Placebo (n=87)
Keller and colleagues*	5	3	1
SKB acute from CSR*	7	3	1
RIAT acute and taper from CSR	11	4 (3 definite, 1 possible)	2 (1 definite, 1 possible)

*Keller and colleagues and CSR mostly reported suicide related events as "emotional lability."

The full details for patients included in this table can be found in appendix 3, along with working notes and directions to where in the CSR the key details can be found. It is possible to take different approaches to moving taper phase events into the continuation phase and reviewing the coding for all cases, especially cases 039, 089, and 106, that were designated suicidal and self injurious behaviours in the RIAT recoding. This would result in different figures.

There were no noteworthy changes in physiological data, which are detailed in appendix F (patient data listings of laboratory tests) in the CSR.

Severity ratings

In the CSR, serious adverse events (defined as an event that "resulted in hospitalization, was associated with suicidal gestures, or was described by the treating physician as serious") were reported in 11 patients in the paroxetine group, five in the imipramine group, and two in the placebo group. Designating an adverse event as serious hinged on the judgment of the clinical investigator. We were therefore unable to make comparable judgments of seriousness, but there are two other methods to approach the issue of severity of adverse events. One is to look at those rated as severe rather than moderate or mild at the time of the event (table 7). A high number and proportion of severe psychiatric events occurred in the paroxetine group. In contrast, few of the

Table 7 | Adverse events (ADECS coded) deemed serious by investigator in Study 329 and reorganised by RIAT analysis to MEDRA system organ class (SOC)

Adverse event (system organ class)	Paroxetine (n=93)	Imipramine (n=95)	Placebo (n=87)
Cardiovascular	1	3	0
Gastrointestinal	25	20	4
Psychiatric	32	4	6
Respiratory	2	1	4
Neurological	7	14	7
Other	3	8	5
Total	70	50	26

doi: 10.1136/bmj.h4320 | *BMJ* 2015;351:h4320 | the bmj

Table 8 | Reasons for withdrawal (86 patients) during acute phase and taper* in Study 329

	Paroxetine (n=93)		Imipramine (n=95)		Placebo (n=87)	
Reason for withdrawal	CSR	RIAT	CSR	RIAT	CSR	RIAT
Adverse event						
Aggression	1	0	0	1	0	0
Mania	1	2	0	0	0	0
Overdose	1	1	0	0	0	0
Depression worsening	0	1	0	0	0	1
Agitation	0	1	0	0	0	0
Suicidality	0	5†	0	2	0	1
Hallucinations	0	0	0	1	0	0
Conduct disorder	1	1	0	0	0	0
Hospital admission/surgery	1	0	1	0	0	0
Fatigue	0	0	1	1	0	0
Sedation	0	1	0	1	0	0
Nausea/vomiting	0	1	2	5	0	1
Rash/acne	0	0	2	3	1	1
Cardiac	0	1	9	15	3	2
Accidental injury	0	0	1	0	0	0
Urinary	0	0	1	1	0	0
Pregnancy	0	0	1	1	0	0
Intercurrent illness‡	6	0	12	0	2	0
Total adverse events (%)	11 (11.8)	14 (15.0)	30 (31.5)	31 (32.6)	6 (6.9)	6 (6.9)
Protocol violation§						
Non-compliance with medication	3	1	4	4	6	4
By investigator	0	0	0	0	0	4
Recreational drug use	0	0	1	1	1	1
Total protocol violations (%)	3 (3.2)	1 (1.1)	5 (5.3)	5 (5.3)	7 (8.0)	9 (10.3)
Other (%)						
Lost to follow-up	5 (5.4)	4 (4.3)	1 (1.1)	1 (1.1)	1 (1.1)	1 (1.1)
Lack of efficacy	3 (3.2)	3 (3.2)	1 (1.1)	0 (0)	6 (6.9)	4 (4.6)
Withdrawn consent	4 (4.3)	5 (5.4)	1 (1.1)	1 (1.1)	1 (1.1)	1 (1.1)
Overall						
Total dropout rate (%)	26 (28)	27 (29)	38 (40)	38 (40)	21 (24)	21 (24)

*Reported in appendix G (tabulations by patient) from CSR and from appendix H CRFs.
†Patient 329.002.00058 was found to have stopped drug three days before attempting suicide. Originally this had been classed as "continuation phase" drop out, but we moved it to "30 day discontinuation" period. Reason for withdrawal was originally "adverse event including intercurrent illness" but we changed it to "suicide attempt." Consequently RIAT analysis found total of 86 withdrawals rather than 85 reported in CSR.
‡We replaced term "adverse event: intercurrent illness" with more specific adverse event terms.
§Four patients enrolled in study violated inclusion criterion. Two had cardiovascular problems, one had C-GAS score >60, and one was "extremely" suicidal at screening. All four were randomised to placebo. It was unclear how to categorise their reasons for discontinuation; we chose "protocol violations."

many cardiovascular events in the imipramine group were rated as severe.

Discontinuations

A second method of approaching the issue of severity of adverse events is to look at rates of discontinuation because of such events. Table 8 shows reasons for withdrawal during the acute phase and taper because of adverse events and other causes. Note that we examined the case report forms from appendix H for all discontinuations reported in appendix G of the CSR. All changes of coding for discontinuation are laid out in table H in appendix 2.

Consideration of the displaced taper in Study 329 revealed a conundrum. In addition to the 86 dropouts from the acute phase noted by SKB, there were 65 dropouts after ratings were completed at week eight. SKB regarded these patients as participants in the continuation phase, although none of them took a continuation phase drug or had a continuation phase rating. The coding for discontinuation was particularly ambiguous for this group.

Most patients stopped at this point were designated by SKB as "lack of efficacy" (table 9). Investigators in four centres reported lack of efficacy as a reason for stopping six patients allocated to placebo even though the HAM-D score was in the responder range and was as low as 2 or 3 points in some instances.

In some cases there were clear protocol violations or factors such as the unavailability of further treatments (placebo in particular). We recategorised the lack of efficacy dropouts based on factors such as adverse events and HAM-D scores. Table 9 shows our analysis of reasons for withdrawal at the end of the acute phase.

Withdrawal effects

The protocol for Study 329 called for a taper phase for all participants and, in addition, a 30 day follow-up period for all those who discontinued because of adverse events. The data in the appendix D of the CSR make it possible to identify adverse events happening in the taper and follow-up periods. These data are presented in table 10.

Effects of other drugs

Table 11 shows data on the effects of other drugs on the adverse events recorded. Patients taking other drugs had more adverse events than those who were not. This effect was slightly more marked in the placebo group, and as such works to the apparent benefit of the active drug treatments in minimising any excess of adverse events over placebo.

Discussion

Principal findings and comparison with original journal publication

Our RIAT analysis of Study 329 showed that neither paroxetine nor high dose imipramine was effective in the treatment of major depression in adolescents, and there was a clinically significant increase in harms with both drugs. This analysis contrasts with both the published conclusions of Keller and colleagues[2] and the way that the outcomes were reported and interpreted in the CSR.

We analysed and reported Study 329 according to the original protocol (with approved amendments). Appendix 1 shows the sources of information we used in preparing this paper, which should help other researchers who want to access the data to check our analysis or to interrogate it in other ways. We draw minimal conclusions regarding efficacy and harms, inviting others to offer their own analysis.

Our re-examination of the data, including a review of 34% of the cases, showed no significant discrepancies in the primary efficacy data. The marked difference between the efficacy outcomes as reported by us and those reported by SKB results from the fact that our analysis kept faith with the protocol's methods and its designation of primary and secondary outcome variables.

The authors/sponsors departed from their study protocol in the CSR itself by performing pairwise

Table 9 | Reasons for withdrawal (65 patients) at end of acute phase according to SKB and RIAT reanalysis in Study 329*

	Paroxetine group (acute completers n=67)		Imipramine group (acute completers n=56)		Placebo group (acute completers n=66)	
Reason for withdrawal	**CSR, appendix G**	**RIAT†**	**CSR, appendix G**	**RIAT†**	**CSR, appendix G**	**RIAT†**
Adverse event						
Aggression/paranoia	1	1	0	0	0	0
Overdose	1	0	0	0	0	0
Depression worsening	0	1	0	0	0	0
Homicidal ideation	0	0	1	1	0	0
Suicidality	0	2	0	0	0	0
Rash	1	1	0	0	0	0
Cardiac	0	0	1	2	0	0
Dry mouth	0	0	0	1	0	0
Total adverse events	3	5	2	4	0	0
Protocol violation						
Non-compliance with study treatment	1	1	2	2	0	0
Recreational drug use	0	0	0	0	1	1
PV by investigator	0	1	0	2	0	3
Total protocol violations	1	2	2	4	1	4
Lost to follow-up						
Total	0	2	0	0	0	0
Lack of efficacy						
Total	9	5	12	8	23	17
Withdrawn consent						
Total	1	1	0	0	4	5
Other						
HAM-D responders	0	1	0	1	0	6
General surgery	1	0	0	0	0	0
No study drugs available	1	0	0	0	3	0
ADHD symptoms	0	0	1	0	0	0
Moved out of state	0	0	0	0	1	0
Total "other" dropouts	2	1	1	1	4	6
Overall total						
Total	16	16	17	17	32	32

HAM-D=Hamilton depression scale, ADHD=attention-deficit/hyperactivity disorder.
*Total discontinued at week 8 (end of acute phase). CSR and paper by Keller and colleagues report 86 patients who withdrew in acute phase, but are silent about these 65 patients who dropped out at end of acute phase.
†After review of reasons for withdrawal from study in the CSR (appendix G), along with review of patient narratives and CRFs where applicable, we proposed changes to these reasons for withdrawal in a proportion of those discontinued.

Table 10 | Adverse events from taper phase of Study 329 according to RIAT (reanalysis study)*

	Paroxetine (n=19)		Imipramine (n=32)		Placebo nN=9)	
System organ class (MedDRA)	**RIAT MedDRA coded**	**Reported as severe**	**RIAT MedDRA coded**	**Reported as severe**	**RIAT MedDRA coded**	**Reported as severe**
Cardiovascular disorders	4	0	9	0	0	0
Gastrointestinal disorders	9	4	18	4	4	0
Psychiatric disorders	15	8	2	0	1	1
Respiratory-thoracic disorders	3	0	1	0	0	0
All other SOCs	16	1	20	5	5	0
Total adverse events	47	13	50	9	10	1

*SKB did not present ADECS analysis for taper phase in clinical study report.

Table 11 | Use of other drugs in month before enrolment, and incidence of adverse events in Study 329

	Paroxetine (n=93)		Imipramine (n=95)		Placebo (n=87)	
	Other drugs	**No other drugs**	**Other drugs**	**No other drugs**	**Other drugs**	**No other drugs**
No (%) of patients	24 (26)	69 (74)	31 (33)	64 (67)	26 (30)	61 (70)
Psychiatric adverse events subgroup* (acute+taper)	15	42	12	21	6	11
Total adverse events (acute+taper)	158	323	220	332	137	193

*Psychiatric adverse events included in this subgroup include: abnormal dreams, aggravated depression, agitation, akathisia, anxiety, depersonalisation, disinhibition, hallucinations, paranoia, psychosis, suicidal ideation/gesture/attempt.

doi: 10.1136/bmj.h4320 | *BMJ* 2015;351:h4320 | the bmj

comparisons of two of the three groups when the omnibus ANOVA showed no significance in either the continuous or dichotomous variables. They also reported four other variables as significant that had not been mentioned in the protocol or its amendments, without any acknowledgment that these measures were introduced post hoc. This contravened provision II of appendix B of the Study 329 protocol ("Administrative Matters"), according to which any change to the study protocol was required to be filed as an amendment/modification.

With regard to adverse events, there were large and clinically meaningful differences between the data as analysed by us, those summarised in the CSR using the ADECS methods, and those reported by Keller and colleagues. These differences arise from inadequate and incomplete entry of data from case report forms to summary data sheets in the CSR, the ADECS coding system used by SKB, and the reporting of these data sheets in Keller and colleagues. SKB reported 338 adverse events with paroxetine and Keller and colleagues reported 265, whereas we identified 481 from our analysis of the CSR, and we found a further 23 that had been missed from the 93 case report forms that we reviewed.

Another reason why the figures of Keller and colleagues are lower than ours is because they presented data only for adverse events reported for 5% of patients or more. For all adverse events combined, their table 3 reported a burden of adverse events with paroxetine 1.2 times that of the burden with placebo. This compares with the figure of 1.4 from our RIAT MedDRA coding of data from the CSR. The figures from CSR and case report forms also differ substantially from other figures quoted by Keller and colleagues because they did not report a category of psychiatric adverse events, but instead grouped such events together with "dizziness" and "headache" under the class "nervous system."

MedDRA distinguishes between neurological and psychiatric system organ classes. We placed headaches in the neurological rather than the psychiatric class. MedDRA allows dizziness to be coded under cardiovascular or neurological classes. Given the dose of imipramine being used, most cases of dizziness seem likely to be cardiovascular, with Keller and colleagues also reporting a high rate of postural hypotension on imipramine. We have thus coded all dizziness under cardiovascular rather than neurological. There is scope for others accessing the data to parse out whether there is sufficient information to code certain instances of dizziness, such as dizziness during paroxetine taper, as neurological, but we have not carried out that more complex analysis.

As reported by Keller and colleagues, dizziness and headache comprised 54 of 115 nervous system events in those taking paroxetine (47%), 83 of 135 events in those taking imipramine (62%), and 50 of 65 events in those taking placebo (77%). The effect of disentangling these two symptoms from psychiatric adverse events unmasks a clinically important difference in psychiatric adverse event profiles between paroxetine and placebo.

There was a major difference between the frequency of suicidal thinking and events reported by Keller and colleagues and the frequency documented in the CSR, as shown in table 6.

With regard to dropouts, Keller and colleagues stated that 69% of patients completed the acute phase. Only 45%, however, went on to the continuation phase, which has not yet been subject to RIAT analysis.

Comparison with other studies

Our findings are consistent with those of other studies, including a recent examination of 142 studies of six psychotropic drugs for which journal articles and clinical trial summaries were both available.[26 27] Most deaths (94/151, 62%) and suicides (8/15, 53%) reported in trial summaries were not reported in journal articles. Only one of nine suicides in olanzapine trials was reported in published papers.

Reporting of adverse events

Our reanalysis of Study 329 showed considerable variations in the way adverse events can be reported, demonstrating several ways in which the analysis and presentation of safety data can influence the apparent safety of a drug. We identified the following potential barriers to accurate reporting of harms (summarised in box 2).

Use of an idiosyncratic coding system

The term "emotional lability," as used in SKB's adverse drug events coding system, masks differences in suicidal behaviour between paroxetine and placebo.

Failure to transcribe all adverse events from clinical record to adverse event database

Our review of case report forms disclosed significant under-recording of adverse events.

Filtering data on adverse events through statistical techniques

Keller and colleagues (and GSK in subsequent correspondence) ignored unfavourable harms data on the grounds that the difference between paroxetine and placebo was not statistically significant, at odds with the SKB protocol that called for primary comparisons to be made using descriptive statistics. In our opinion, statistically significant or not, all relevant primary and secondary outcomes and harms outcomes should be explicitly reported. Testing for statistical significance is most appropriately undertaken for the primary

Box 2 Potential barriers to accurate reporting of harms

- Use of an idiosyncratic coding system
- Failure to transcribe all adverse events from clinical record to adverse event database
- Filtering data on adverse events through statistical techniques
- Restriction of reporting to events that occurred above a given frequency in any one group
- Coding event under different headings for different patients (dilution)
- Grouping of adverse events
- Insufficient consideration of severity
- Coding of relatedness to study medication
- Masking effects of concomitant drugs
- Ignoring effects of drug withdrawal

outcome measures as study power is based on these. We have not undertaken statistical tests for harms as we know of no valid way of interpreting them. To get away from a dichotomous (significant/non-significant) presentation of evidence, we opted to present all original and recoded evidence to allow readers their own interpretation. The data presented in appendix 2 and related worksheets lodged at www.Study329.org will, however, readily permit other approaches to data analysis for those interested, and we welcome other analyses.

Restriction of reporting to events that occurred above a given frequency in any one group

In the paper by Keller and colleagues, reporting only adverse events that occurred in more than 5% of patients obscured the harms burden. In contrast, we report all adverse events that have been recorded. These are available in table E in appendix 2.

Coding event under different headings for different patients (dilution)

The effect of reporting only adverse events that have a frequency of more than 5% is compounded when, for instance, agitation might be coded under agitation, anxiety, nervousness, hyperkinesis, and emotional lability; thus, a problem occurring at a rate of >10% could vanish by being coded under different subheadings such that none of these reach a threshold rate of 5%.

Aside from making all the data available so that others can scrutinise it, one way to compensate for this possibility is to present all the data in broader system organ class groups. MedDRA offers the following higher levels: psychiatric, cardiovascular, gastrointestinal, respiratory, and other. In table E in appendix 2, the adverse events coded here under "other" are broken down under the additional MedDRA SOC headings, including general, nervous system, metabolic, and pregnancy.

Grouping of adverse events

Even when they are presented in broader system groups, grouping common and benign symptoms with more important ones can mask safety issues. For example, in the paper by Keller and colleagues, common adverse events such as dizziness and headaches are grouped with psychiatric adverse events in the "nervous system" SOC heading. As these adverse events are common across treatment arms, this grouping has the effect of diluting the difference in psychiatric side effects between paroxetine, imipramine, and placebo.

We followed MedDRA in reporting dizziness under "cardiovascular" events and headache under "nervous system." There might be better categorisations; our grouping is provisional rather than strategic. In table E in appendix 2, we have listed all events coded under each system organ class heading, and we invite others to further explore these issues, including alternative higher level categorisation of these adverse events.

Insufficient consideration of severity

In addition to coding adverse events, investigators rate them for severity. If no attempt is made to take severity into account and include it in reporting, readers could get the impression that there was an equal burden of adverse events in each arm, when in fact all events in one arm might be severe and enduring while those in the other might be mild and transient.

One way to manage this is to look specifically at those patients who drop out of the study because of adverse events. Another method is to report those adverse events coded as severe for each drug group separately from those coded as mild or moderate. We used both approaches (see tables 7 and 8).

Coding of relatedness to study medication

Judgments by investigators as to whether an adverse event is related to the drug can lead to discounting the importance of an effect. We have included these judgments in the worksheets lodged at www.Study329.org, but we have not analysed them because it became clear that the blinding had been broken in several cases before relatedness was adjudicated by the original investigators and because some judgments were implausible. For instance, it is documented on page 279 in the CSR that an investigator, knowing the patient was on placebo, declared that a suicidal event was "definitely related to treatment" on the grounds that "the worsening of depression and suicidal thought were life threatening and definitely related to study medication [known to be placebo] in that there was a lack of effect." Notably, of the 11 patients with serious adverse events on paroxetine (compared with two on placebo) reported in the paper by Keller and colleagues, only one "was considered by the treating investigator to be related to paroxetine treatment," thus dismissing the clinically important difference between the paroxetine and placebo groups for serious adverse events.

Masking effects of concomitant drugs

In almost all trials, patients will be taking concomitant drugs. The adverse events from these other drugs will tend to obscure differences between active drug treatment and placebo. This might be an important factor in trials of treatments such as statins, where patients are often taking multiple drugs.

Accordingly, we also compared the incidence of adverse events in patients taking concomitant drugs with the incidence in those not taking other drugs. Other drugs were instituted in the course of the study that we have not analysed, but the data are available in tables K and L in appendix 2 and worksheets lodged at www.Study329.org and in appendix B from the CSR. There are several other angles in the data available at www.Study329.org that could be further explored, such as the effects of withdrawal of concomitant drugs on adverse event profiles, as the spreadsheets document the day of onset of adverse events and the dates of starting or stopping any concomitant drugs. Another option to explore is the possibility of any

doi: 10.1136/bmj.h4320 | *BMJ* 2015;351:h4320 | the bmj

prescribing cascades triggered by adverse events related to study drugs.

Ignoring effects of drug withdrawal

The protocol included a taper phase lasting 7-17 days that investigators were encouraged to adhere to, even in patients who discontinued because of adverse events. The original paper did not analyse these data separately. The increased rates of psychiatric adverse events that emerged during the discontinuation phase in our analysis are consistent with dependence on and withdrawal from paroxetine, as reported by Fava.[29]

RIAT process

This RIAT exercise proved to be extremely demanding of resources. We have logged over 250 000 words of email correspondence among the team over two years. The single screen remote desktop interface (that we called the "periscope") proved to be an enormous challenge. The efficacy analysis required that multiple spreadsheet tables were open simultaneously, with much copying, pasting, and cross checking, and the space was highly restrictive. Gaining access to the case report forms required extensive correspondence with GSK.[12] Although GSK ultimately provided case report forms, they were even harder to manage, given that we could see only one page at a time. It required about a thousand hours to examine only a third of the case report forms. Being unable to print them was a considerable handicap. There were no means to prepare packets for multiple independent coders, to decrease bias; to make annotations or use margin comments; or to sort and collate the adverse event reports. Our experience highlights that hard copies as well as electronic copies are crucial for an enterprise like this.

Our analysis indicates that although CSRs are useful, and in this case all that was needed to reanalyse efficacy, analysis of adverse events requires access to individual patient level data in case report forms.

Because we have been breaking new ground, we have not had precedents to call on in analysis and reporting. We await with interest other efforts to do something similar.

Strengths and limitations of this study

Study 329 was a randomised controlled trial with a reasonable sample size. There was, however, evidence of protocol violations, including some cases of breaking of blinding. The coding of adverse events by the original investigators raised the possibility that some other data might be unreliable.

The trial lasted for only eight weeks. Participants had relatively chronic depression (mean duration more than one year), which would limit the generalisability of the results, particularly in primary care, because many cases of adolescent depression have shorter durations.[28] Generalisability to primary care would also be limited by the fact that participants were recruited through tertiary settings.

The RIAT analysis broke new ground but was limited in that we could check only 34% (93/275) of case report forms. Time and resources prevented access to all forms because of the difficulties in using the portal for accessing the study data and because considerable amounts of data were missing.

The analysis generated a useful taxonomy of potential barriers to accurate reporting of adverse events and, even allowing for the above limitations, showed the value of permitting access to data.

Conclusion and implications for research and policy

Contrary to the original report by Keller and colleagues, our reanalysis of Study 329 showed no advantage of paroxetine or imipramine over placebo in adolescents with symptoms of depression on any of the prespecified variables. The extent of the clinically significant increases in adverse events in the paroxetine and imipramine arms, including serious, severe, and suicide related adverse events, became apparent only when the data were made available for reanalysis. Researchers and clinicians should recognise the potential biases in published research, including the potential barriers to accurate reporting of harms that we have identified. Regulatory authorities should mandate accessibility of data and protocols.

As with most scientific papers, Keller and colleagues convey an impression that "the data have spoken." This authoritative stance is possible only in the absence of access to the data. When the data become accessible to others, it becomes clear that scientific authorship is provisional rather than authoritative.

We thank Carys Hogan for database work and Tom Jefferson and Leemon McHenry for comments on earlier drafts.

The SmithKline Beecham study was registered as No 29060/329. The protocol was SmithKline Beecham study 29060/329, final clinical report (acute phase), appendix a, Protocol, from p 531.[13] The study was funded by SmithKline Beecham. The data analysis protocol for RIAT reanalysis was submitted to GSK on 28 October 2013 and approved by GSK on 4 December 2013.

Contributors: Conception/design of the work: DH, JJ, JMN. Acquisition of data: JJ (negotiation with GSK); CT and EA-J (RIATAR); JMN (efficacy data using GSK online remote system); JLN (harms data using GSK online remote system). Data analysis: JMN (efficacy); JLN and DH (harms). Data interpretation: all authors. Drafting the work and revising it critically for important intellectual content, final approval of the version to be published: all authors. All authors agree to be accountable for all aspects of the work. JJ is guarantor. The first four authors made equal contribution to the paper.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: DH has been and is an expert witness for plaintiffs in legal cases involving GlaxoSmithKline's drug paroxetine. He is also a witness for plaintiffs in actions involving other antidepressants with the same mechanism of action as paroxetine, and is on the advisory board of the Foundation for Excellence in Mental Health Care. DH and JLN are founder members of RxISK. JJ has been paid by Baum, Hedlund, Aristei and Goldman, Los Angeles, CA, to provide expert analysis and opinion about documents obtained from GlaxoSmithKline in a class action over Study 329, and from Forest in relation to paediatric citalopram randomised controlled trials. Some of the authors are in discussions with an academic publisher regarding adapting the case of Study 329 as a book for educational purposes.

Ethical approval: The protocol and statement of informed consent were approved by an institutional review board before each centre's initiation, in compliance with 21 United States Code of Federal

Regulations (CFR) Part 56. Written informed consent was obtained from each patient before entry into the study, in compliance with 21 CFR Part 50. Case report forms were provided for each patient's data to be recorded (Final Clinical Report page 000030). The sample informed consent is provided in the appendix to the protocol, appendix C, pp 000590-4. No further information is available regarding the particular institutional review board that approved the study.

Transparency: JJ affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Data sharing: Clinical study reports, detailed data tables, and programming code are available on the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.bv8j6) and at www.Study329.org/

1 Doshi P, Dickersin K, Healy D, Vedula SS, Jefferson T. Restoring invisible and abandoned trials: a call for people to publish the findings. *BMJ* 2013;346:f2865.
2 Keller MB, Ryan ND, Strober M, et al. Efficacy of paroxetine in the treatment of adolescent major depression: a randomized, controlled trial. *J Am Acad Child Adolesc Psychiatry* 2001;40:762-72.
3 McHenry L, Jureidini J. Industry-sponsored ghostwriting in clinical trial reporting: a case study. *Account Res* 2008;15:152-67.
4 Jureidini J, McHenry L, Mansfield P. Clinical trials and drug promotion: selective reporting of study 329. *Int J Risk Saf Med* 2008;20:73-81.
5 Jureidini J, McHenry L. Conflicted medical journals and the failure of trust. *Account Res* 2011;18:45-54.
6 Kraus JE, letter to Jon Jureidini. 2013. www.bmj.com/content/suppl/2013/11/12/bmj.f6754.DC1/doshinov16.ww1_default.pdf.
7 Treasure T, Monson K, Fiorentino F, Russell C. The CEA Second-Look Trial: a randomised controlled trial of carcinoembryonic antigen prompted reoperation for recurrent colorectal cancer. *BMJ Open* 2014 May 13;4:e004385.
8 Ebrahim S, Sohani ZN, Montoya L, et al. Reanalyses of randomized clinical trial data. *JAMA* 2014;312:1024-32.
9 SmithKline Beecham. A multi-center, double-blind, placebo controlled study of paroxetine and imipramine in adolescents with unipolar major depression –acute phase, Final clinical report. www.gsk.com/media/389566/depression_329_full.pdf.
10 Healthy Skepticism International News. Paxil Study 329: paroxetine vs imipramine vs placebo in adolescents. 2010. www.healthyskepticism.org/global/news/int/hsin2010-01.
11 SAS Solutions OnDemand. www.ondemand.sas.com/sam/?sid=1393012805&rid=DqqHX0rCqAWZ7TJICIPijRTscQ.
12 Correspondence between Jureidini and GSK. Rapid responses to putting GlaxoSmithKline to the test over paroxetine. *BMJ* 2013;347:f6754. www.bmj.com/content/347/bmj.f6754/rapid-responses
13 SmithKline Beecham. A multi-center, double-blind, placebo controlled study of paroxetine and imipramine in adolescents with unipolar major depression 1993/amended 1996. www.gsk.com/media/360485/329-AppA.PDF.
14 Diagnostic and statistical manual of mental disorders, third edition, revised (DSM-III-R). American Psychiatric Association, 1987.
15 Fawcett J, Epstein P, Fiester SJ, Elkin I, Autry JH. Clinical management—imipramine/placebo administration manual. NIMH Treatment of Depression Collaborative Research Program. *Psychopharmacol Bull* 1987;23:309-24.
16 Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* 1967;6:278-96.
17 Sigafoos AD, Feinstein CB, Damond M, Reiss D. The measurement of behavioral autonomy in adolescence: the Autonomous Functioning Checklist. *Adolesc Psychiatry* 1988;15:432-62.
18 SKB. Draft Minutes: 4/22/97 Teleconference. Paroxetine Study 329 efficacy analysis. www.healthyskepticism.org/files/docs/gsk/paroxetine/study329/970422teleconference.pdf.
19 GlaxoSmithKline, Paroxetine—paediatric and adolescent patients. www.gsk.com/en-gb/media/resource-centre/paroxetine/paroxetine-paediatric-and-adolescent-patients/.
20 Winter C. MedDRA in clinical trials—industry perspective SFDA- ICH MedDRA Workshop, Beijing, 13- 14 May 2011. www.meddra.org/sites/default/files/page/documents_insert/christina_winter_2_meddra_in_clinical_trials_industry_perspective.pdf.
21 Jureidini JN, Nardo JM. Inadequacy of remote desktop interface for independent reanalysis of data from drug trials. *BMJ* 2014;349:g4353.
22 Fitzgerald K, Healy D. Dystonias and dyskinesias of the jaw associated with the use of SSRIs. *Human Psychopharmacol* 1995;10:215-20.
23 Kline RB. Beyond significance testing. Statistics reform in the behavioral sciences. 2nd ed. American Psychological Association, 2013.
24 R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. www.R-project.org/.
25 Brecher M. Review and evaluation of clinical data. Original NDA 20—031. Paroxetine (Aropax). Efficacy review. SmithKline Beecham Pharmaceuticals, 1991.
26 Hughes S, Cohen, D, Jaggi R. Differences in reporting serious adverse events in industry sponsored clinical trial registries and journal articles on antidepressant and antipsychotic drugs: a cross-sectional study. *BMJ Open* 2014;4:e005535.
27 Maund E, Tendal B, Hróbjartsson A, et al. Benefits and harms in clinical trials of duloxetine for treatment of major depressive disorder: comparison of clinical study reports, trial registries, and publications. *BMJ* 2014;348:g3510
28 Lewinsohn PM, Clarke GN, Seeley JR, Rohde P. Major depression in community adolescents: age at onset, episode duration, and time to recurrence. *J Am Acad Child Adolesc Psychiatry* 1994;33:809-18.
29 Fava M. Prospective studies of adverse events related to antidepressant discontinuation. *J Clin Psychiatry* 2006;67(suppl 4):14-21.

Appendix 1: RIAT audit record
Appendix 2: Supplementary tables A-M
Appendix 3: Supplementary information on suicidal and self-injurious behaviours in Study 329